DOCUMENT RESUME

ED 298 155 TM 012 252

AUTHOR Robey, Randall R.; Barcikowski, Robert S.
TITLE The Number of Iterations in Monte Carlo Studies of Robustness.
PUB DATE Apr 88
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; *Monte Carlo Methods; Simulation; *Statistical Significance
IDENTIFIERS Confidence Intervals (Statistics); Iterative Methods; *Robustness; *Type I Errors

ABSTRACT

A recent survey of simulation studies concluded that an overwhelming majority of papers do not report a rationale for the number of iterations carried out in Monte Carlo robustness (MCR) experiments. The survey suggested that researchers might benefit from adopting a hypothesis testing strategy in the planning and reporting of simulation studies. This paper presents a table of the number of iterations necessary to detect departures from a series of nominal Type I error rates based upon hypothesis testing logic. The table is indexed by effect size, by significance level, and by power level for the two-tailed test that a proportion equals some constant. An alternative approach based upon the construction of a confidence interval is discussed and dismissed. The MCR research design demands an adequate definition of robustness and a sufficient sample size to detect departures from that definition. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298155

The Number of Iterations in Monte Carlo Studies of Robustness

Randall R. Robey
Southern Illinois University

Robert S. Barcikowski
Ohio University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RANDALL R. ROBEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of the
American Educational Research Association,
New Orleans, April, 1988.

TM 012 252

BEST COPY AVAILABLE

Abstract

A recent survey of simulation studies concluded that an overwhelming majority of papers do not report a rationale for the number of iterations carried out in this type of experiment. The survey authors suggested that researchers might benefit from adopting a hypothesis testing strategy in the planning and reporting of simulation studies.

This paper presents a table of the number of iterations necessary to detect departures from a series of nominal Type I error rates based upon hypothesis testing logic. The table is indexed by effect size, by significance level, and by power level for the two-tailed test that a proportion equals some constant. An alternate approach based upon the construction of a confidence interval is discussed and dismissed.

The Number of Iterations in Monte Carlo Studies of Robustness

Introduction

The Monte Carlo robustnes experiment is frequently used to estimate the Type I error characteristics of one or more algorithms under various assumption violation conditions (Hoagalin and Andrews, 1975). In this type of experiment, the comparison of interest is between the nominal Type I error rate (α) and the actual Type I error rate (π). The estimate of the actual Type I error rate ($\hat{\pi}$) is the observed proportion of the total number (n) of calculated test statistics exceeding a critical test size under the null hypothesis (Olson, 1973). If it can be reasoned that π approximates α within an acceptable tolerance, the performance of the algorithm is said to be robust with respect to the specified violation condition.[1]

For example, let's assume that an investigator is interested in estimating the Type I error perfomance of an F test under some assumption violation condition. Further, for the sake of simplicity, let's assume that the investigator is interested in the Type I error performance of the test at the .05 level only. To complete the example, let's say the investigator intends to carry out the calculation of the F statistic for each of 1000 samples from a population where the null hypothesis is true and where the violation condition exists. In this situation, then, .05 is the Type I error rate (α). The unknown proportion of incorrect rejections for the entire population based on some .05 critical value, represented by π , is the actual Type I error rate. The estimate $\hat{\pi}$ is the total number of incorrect rejections observed in the 1000 samples based upon the same critical value.

Among other things, when constructing the research design for a Monte Carlo robustness experiment, a researcher must address two related issues. First, how large must the discrepancy between α and π become before the test ceases to be robust. Second, what number of iterations must be carried out in the simulation procedure in order that the performance of the algorithm can be interpreted with confidence?

Bradley (1978) examined the first issue and proposed guidelines for developing a definition of robustness. These guidelines are reviewed in the following section. Then, based in part upon Bradley's guidelines, we describe a method for

---

[1] For other definitions of the term 'statisical robustness', see Huber (1981).

determining the necessary number of iterations in Monte Carlo studies of robustness. Moreover, we compare the recommended strategy to a competing strategy, i.e, the confidence interval approach. In a subsequent section, we relate this method to some reports of Monte Carlo robustness experiments which are found in contemporary literature.

Defining Robustness

Concerning the first research design decision mentioned above, Bradley (1978) observed that many practioners may be unreasonably generous in defining robustness. Bradley suggested defining robustness as an interval about α, the bounds of which are proportional to α. Bradley (1978, p. 146) defined a "fairly stringent criterion" as $.9\ \alpha \leq \alpha \leq 1.1\ \alpha$, which can be written as $\alpha \pm 1/10\ \alpha$, and a "liberal criterion" as $.5\ \alpha \leq \alpha \leq 1.5\ \alpha$, which can be written as $\alpha \pm 1/2\ \alpha$.

In this paper, these intervals plus two additional robustness intervals are reported in order to address a range of interests. The additions are an intermediate criterion given by $\alpha \pm 1/4\ \alpha$, and a fairly liberal criterion given by $\alpha \pm 3/4\ \alpha$. Based upon these four robustness intervals, the remainder of this paper is designed to provide researchers with the information necessary to objectively select the number of iterations to be carried out in Monte Carlo investigations of Type I error.

Selecting the Number of Iterations

Hauck and Anderson (1984), in a survey of simulation studies, found that only nine percent of the surveyed reports included a justification for the number of iterations utilized. Further, Hauck and Anderson (1984, p. 125) reported that

> "No paper indicated consideration of the power to detect a difference from some null value, as would be appropriate for checking the level of a significance testing procedure or coverage probability of a confidence interval method."

The decision making process described below rests squarely upon hypothesis testing logic, the same approach found lacking in the reports surveyed by Hauck and Anderson (1984). This procedure is motivated by calls for increased scientific rigor in the design and reporting of Monte Carlo experiments (Halperin, 1976; Hauck and Anderson, 1984; Hoaglin and Andrews, 1975).

The Hypothesis Testing Appraoch

The general form of the non-directional null hypothesis that a population proportion equals some constant (c) is written as

$$H_o: \pi = c$$

In Monte Carlo robustness experiments, this null hypothesis can be evaluated using a two-tailed proportions test where $c = \alpha$. Since the power of this test to detect departures of π from α is not the same when $\pi < \alpha$ as it is when $\pi > \alpha$, comparisons of π and α are facilitated by transforming each by

$$\phi_x = 2 \arcsin \sqrt{x} \qquad (1)$$

where arcsin is given in radians and x is a proportion. The value of $| \phi_\alpha - \phi_{\hat{\pi}} |$ is tested against the following critical value

$$Z_{1-\omega/2} \sqrt{2/n} \sqrt{.5} \qquad (2)$$

where Z is the standard unit normal deviate, and where ω is the Type I error rate for the proportions test (Cohen, 1977, pp. 460 and 212).

Once the Type I error rate and the power level for the proportions test have been selected, the number of iterations (n) is given by Cohen (1977, p. 461) as

$$n = 2 \left[\frac{Z_{1-\omega/2} + Z_{1-\beta}}{|\phi_{\alpha'} - \phi_\alpha| \sqrt{2}} \right]^2 \qquad (3)$$

where α' is the upper bound of the robustness interval and β represents Type II error. The upper bound is used in the calculation of n since the arcsin transformation causes the interval about ϕ_α to be asymmetric where the distance from ϕ_α to the $\phi_{\alpha'}$ is less than the distance from ϕ_α to the transformed lower bound. As a result, departures from α toward a proportion of .5 are harder to detect and, therefore, require a few more observations relative to the number necessary to detect departures from α which are further out in the tail. Some investigators (e.g., Tomarken and Serlin, 1986) may not be interested in detecting conservative departures from α. In this

case, n can be calculated for a one-tailed test.

Tabled Values of n. Table 1 contains the number of iterations necessary to detect departures from α for each of the four definitions of robustness via the two-tailed proportions test. In this table, the values for nominal alpha include .10, .05, and .01. For each nominal alpha, the entries are indexed by three Type I error rates for the two-tailed proportions test, ω, (i.e., ω = .01, .05, and .10), and by three statistical powers for the two-tailed proportions test, $1 - \beta$, (i.e., $1 - \beta$ = .7, .8, and .9).

Examination of the tables reveals that when using the hypothesis testing approach to select n, simply carrying out 1000 iterations is appropriate only for the larger values of α in combination with the more liberal definitions of robustness.

The adoption of a more stringent definition of robustness when attempting to detect departures from a small value of α, say .01 or less, requires an n which is substantially larger than that found in most Monte Carlo studies. For example, 121312 iterations are required to detect departures from Bradley's most stringent definition of robustness for α at .01, when the power for the proportions test is set at .80, and ω is set at .01.

An Alternate Procedure: The Confidence Interval

Several researchers choose to follow the advice of Glass, Peckham and Sanders (1972) to evaluate the outcome of robustness experiments. Glass et al. recommended against attaching meaning to departures of $\hat{\pi}$ from α which are less than approximately two standard errors in magnitude.

This practice amounts to selecting some large value for n ($n \geq 500$) and then solving for a critical bandwidth which defines robustness. However, the argument presented here is that it seems much more prudent to first set the bandwidth based upon an acceptable definition of robustness and then to solve for n.

A Comparison. The hypothesis testing approach, and the two standard error confidence interval approach, for determining the number of iterations to be carried out in a Monte Carlo investigation of robustness represent different perspectives on the problem of interpreting the results obtained in this type of experiment. Both of these analyses comment on the approximation of π to α. However, meaningful differences distinguish the two procedures.

By statistical inference, the hypothesis testing approach comments on the approximation of π to α by direct comparison.

Here, the emphasis is placed on detecting the non-null case (e.g., $\pi \neq \alpha$) with known Type I and Type II error rates.

Alternatively, the confidence interval approach comments on the accuracy with which $\hat{\pi}$ estimates π. This method considers only Type I error in the definition of confidence. The inference to α occurs only when the researcher examines the confidence interval for the presence of α. Because this inference does not occur with a known Type II error rate, the confidence interval approach does not compare favorably to the hypothesis testing approach. As a result, we conclude that the hypothesis testing approach provides the best analysis alternative for examining robustness in Monte Carlo experiments.

Literature Examples

Consider Maxwell and Bray (1986) who examined the robustness of the quasi F statistic under departures from the sphericity assumption using the liberal criterion $\alpha \pm 1/2\ \alpha$ for $\alpha = .05$. Maxwell and Bray report the results of an inferential test on their data, however, the exact nature of that test is not clear. Nevertheless, had Maxwell and Bray used the two-tailed proportions test on their 1000 iterations with ω set at .05, the resulting statistical power would have exceeded .90 (see Table 1).

In another Monte Carlo study, Koch and Yang (1986) used 1000 iterations to examine the Type I and Type II error performances of an asymptotic test which they derived to evaluate the independence of two time series. It should be noted that Koch and Yang did not investigate an assumption violation condition. Rather, their Type I error results were used to estimate that particular property of the asymptotic test. However, the decision making process regarding the acceptability of Type I error performance is essentially the same as in a robustness study.

Koch and Yang (1986) chose to employ the two standard error method to examine their results for departures from α at .10, .05 and at .01. As a result, their robustness tolerances were: $.10 \pm .019$ for $\alpha = .10$, $.05 \pm .014$ for $\alpha = .05$, and $.01 \pm .006$ for $\alpha = .01$. In order to maintain the statistical power of the proportions test at .80, it can be determined from Equation 3 that 2123, 2210, and 2536 iterations would be necessary for α at .10, .05, and .01, respectively. However, when $n = 1000$ and $\omega = .05$, the statistical powers of the two-tailed proportions test to detect effects of these magnitudes are: .49 for $\alpha = .10$, .47 for $\alpha = .05$, and .42 for $\alpha = .01$. From a statistical inference perspective, then, it can be seen that the interpretation of departures of $\hat{\pi}$ from α based on 1000 iterations

is a risky endeavor vis a vis Type II error. That is, a substantial threat to the statistical conclusion validity of the experiment (see Cook and Campbell, 1979) might very well bias the interpretation of the observations.

Conclusion

The most difficult and subjective decision in the recommended procedure resides in the selection of an effect size. That is, deciding how large the difference between π and α must become before the algorithm under investigation will be described as non-robust. It seems reasonable that the appropriate effect size value for any analysis should vary with the stringency required in the application of the algorithm. Moreover, when considering the relative values and costs of this difficult decision, particular attention should be paid to the fact that small changes in the fraction of α used in definition robustness can have a very large impact on the required number of iterations.

In conclusion, consider the following. The Monte Carlo robustness procedure is used to estimate the distributional properties of an algorithm under conditions where these properties cannot be empirically derived. Moreover, these 'conditions' often characterize the data which behavioral scientists more than occasionally find interesting. As a result, educational researchers often find it necessary to conduct Monte Carlo experiments in an effort to improve the application scientist's ability to answer those research questions which may not best answered by the blind administration of some standard analysis. It follows that these application scientists, then, must often rely upon the interpretation of robustness results in order to best analyze their data. As a result, the interpretation of Monte Carlo results can affect a considerable body of subsequent literature. The implication of the situation is quite clear. The Monte Carlo robustness research design demands an adequate definition of robustness and a sufficient sample size to detect departures from that definition.

REFERENCES

Bradley, J. V. (1978), "Robustness?," British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Cohen, J. (1977), Statistical power analysis for the behavioral sciences, (2nd ed.). New York: Academic Press.

Cook, T. D., & Campbell, D. T. (1979), Quasi-experimentation: Design & analysis for field settings. Boston: Houghton Mifflin Co.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972), "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance," Review of Educational Research, 42, 237-288.

Halperin, S. (1976), "Design of Monte Carlo Studies," A paper presented before the annual convention of the American Educational Research Association, San Francisco.

Hauck, W. W., & Anderson, S. (1984), "A survey regarding the reporting of simulation studies," The American Statistician, 38, 214-216.

Hoaglin, D. C., & Andrews, D. F. (1975), "The reporting of computation-based results in statistics," The American Statistician, 29, 122-126.

Huber, P. J. (1981), Robust statistics. New York: Wiley.

Koch, P. D., & Yang, S. S. (1986), "A method for testing the independence of two time series that account for a potential pattern in the cross-correlation function," Journal of the American Statistical Association, 81, 533-544.

Maxwell, S. E., & Bray, J. H. (1986), "Robustness of the quasi F statistic to violations of sphericity," Psychological Bulletin, 99, 416-421.

Olson, C. L. (1973), A Monte Carlo Investigation of the Robustness of Multivariate Analysis of Variance. Unpublished doctoral dissertation, University of Toronto.

Tomarken, A. J., & Serlin, R. C. (1986), "Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures," Psychological Bulletin, 99, 90-99.

Table 1. Iterations Necessary to Detect Departures of π from α Using the Two-tailed Proportions Test.

			Magnitude of Departure			
α	$1-\beta$	ω	$\alpha \pm 1/10\alpha$	$\alpha \pm 1/4\alpha$	$\alpha \pm 1/2\alpha$	$\alpha \pm 3/4\alpha$
		.10	4419	750	204	94
	.7	.05	5796	983	268	128
		.01	9027	1531	417	200
		.10	5810	986	269	129
.10	.8	.05	7375	1251	341	163
		.01	10973	1861	507	243
		.10	8047	1365	372	178
	.9	.05	9873	1675	456	218
		.01	13980	2371	846	309
		.10	9356	1594	437	211
	.7	.05	12271	2091	573	276
		.01	19111	3256	893	430
		.10	12301	2096	575	277
.05	.8	.05	15614	2660	729	351
		.01	23233	3958	1085	523
		.10	17038	2903	796	383
	.9	.05	20902	3561	976	470
		.01	29600	5042	1382	666
		.10	48852	8348	2300	1113
	.7	.05	64072	10948	3017	1460
		.01	99790	17051	4678	2274
		.10	64227	10975	3024	1464
.01	.8	.05	81527	13931	3838	1858
		.01	121312	20729	5711	2764
		.10	88963	15201	4188	2027
	.9	.05	109141	18649	5138	2487
		.01	154556	26409	7276	3521

NOTE. The values of ω and $1-\beta$ are a priori Type I error rates and power levels of the two-tailed proportions test. The values of α are nominal alpha levels.